This Paper Was Written with the Help of ChatGPT: Exploring the Consequences of AI-Driven Academic Writing on Scholarly Practices

Hongming Li
University of Florida
hli3@ufl.edu

Seiyon Lee
University of Florida
leeseiyon@ufl.edu

Anthony F. Botelho
University of Florida
abotelho@coe.ufl.edu

ABSTRACT

Recent advances in the development of large language models (LLMs) have led to power innovative suites of generative AI tools that are capable of not only simulating human-like-dialogue but also composing more complex artifacts, such as social media posts, essays, and even research articles. While this abstract has been written entirely by a human without any input, consultation, or revision from a generative language model, it would be difficult for a reader to discern the difference. Aside from some notable risks, questions remains as to how we should consider the originality of human work that are influenced or partially refined by a generative language model. We present this paper as both a case study into the usage of generative models to support the writing of academic papers but also as an example of how open science practices can help address several issues that have been raised in other contexts and communities. This paper neither attempts to promote nor contest the use of these language models in any writing task. The goal of our work to provide insight and guidance into the ethical and effective usage of these models within this domain.

Keywords

ChatGPT, AI Detectors, Large Language Model, Generative AI, Artificial Intelligence, Academic Writing, Open Science Framework

1. INTRODUCTION

Recent emergence of notable large language models (LLMs) as OpenAI's ChatGPT marks a pivotal shift not only in the domain of teaching and learning but also in the broader scientific communities. The integration of such generative AI tools into the writing process of research offers novel possibilities but it also raises ethical and philosophical questions around the depth of their utility. It is particularly concerning with the comparative lack of sophistication in writing aids (e.g. Grammarly [17], Quillbot [18], and even simple spell checking. Worse than the evidence of"unexpected brittleness" [22] is that the black box nature makes it possible for errors to easily go unnoticed. Furthermore, the question of originality and intellectual contribution must be considered when these models play a role in the creative process.

While some have explored the policies surrounding the inclusion of tools such as ChatGPT as co-authors on scholarly articles [35, 8], the purpose of this paper is not to argue for or against such cases. We will acknowledge upfront that a large language model, specifically GPT-4 Turbo, was used to help write and revise portions of this paper. Rather than take a stance on what we, as a scientific community *should* allow, this paper approaches this topic from the perspective that many researchers *will* be using this and similar tools to aid in various aspects of writing and conceptualization processes. This may either be met with policing action, perhaps by using AI-writing detection tools, or left to authors to decide how to acknowledge how these models may have been used.

The purpose of this paper is to examine and discuss the potential risks of using AI-detection tools to identify works that may have been written with the help of LLMs. Recent studies (e.g. [19]) have begun to scrutinize the effectiveness of current AI detectors in distinguishing between human and AI-generated texts, and presented the complexity of this issue as limitations. Instead, we further offer guidance as to how concerns pertaining to the use of these tools may be addressed through existing open science practices. This paper seeks to contribute to this ongoing discourse by exploring the use of generative models like ChatGPT in the composition of academic research articles. Specifically, we address the pressing need for the academic publishing community to adapt and evolve in response to the rise of generative AI tools, emphasizing the importance of transparency, ethical considerations, and the pursuit of best practices. In our case study, we examine the practical, ethical, and methodological implications of leveraging such technologies in scholarly writing with following research questions: 1. How accurately can current AI content detectors identify AI-generated text from LLMs like GPT-4 Turbo? 2. What adjustments do the academic publishing community need to make to promote transparency and explore best practices in response to the rise of generative AI tools such as ChatGPT?

2. BACKGROUND

The emergence of generative AI has led to transformations across domains and contexts. Within a week of its debut in

November 30, 2022, OpenAI's ChatGPT attracted millions of users with its capacity to simulate a natural conversation with human users. Historically, chatbots were first conceived as a commitment to create machines which emulated and exhibited human-like behavior [36], and ALICE earned honorary recognition as the first chatbot worthy of winning the annual Turing Test award [38]. While chatbots have also attained commercial success (e.g., Apple's Siri, IMB's Watson, and Google's Assistant), the release of GPT-3 (Generative Pre-Trained Transformer) marked a paradigm shift. As the third iteration of the GPT series, it is characterized by great scale of learned model parameters (175 billion parameters), and use of an improved fine-tuning method [12] over previous iterations and other similar models. Some of the documented capabilities span not only creative writing (e.g., poems, stories, and essays) and academic writing [37], but also more technical uses such as debugging code [23, 39].

Generative AI has also been recognized as boosting efficiency in some of the resource-intensive stages of scientific discovery [30]. Some of the common usages include literature review, revising a draft, and editing a manuscript. A recent study explored the possibility of using LLMs for conducting literature review, and identified the major benefits as improved efficiency, more comprehensive coverage, and scaffolds for writing process [7]. Other cases on the use of Generative AI also have been reported, such as analyzing textual data [28], generating summaries [31], and elaborating on technical language [34]. In the field of education, researchers are using AI as a solution to various challenges in traditional learning processes, optimizing learners' experiences, and providing support for educators [9].

With a wider range of tools to accelerate research, it is now unclear how we can best draw the line between uniquely human enterprise and under the influence of AI-enabled assistance. The argument against AI's authorship is more concerned with the issue of accountability [16, 26]. Some leading venues demand the authors to be explicit with their use of AI in the authoring process (e.g. AIED[1], ACL[2]). The policy clearly states that AI does not satisfy the criteria for authorship, as they cannot held accountable for the work. Some voices across different disciplines have converged toward the need to establish a code of practice and ensure that ChatGPT and other Generative AI solutions uphold high standards of ethics in research [16]. The approach is to capitalize on the AI-enabled productivity boost but to maintain the core of human oversight. Hence, AI is to be considered as a tool to help augment human intelligence and contribute to human-led efforts at best [16, 15].

To ensure an ethical and responsible use of Generative AI, the development of reliable AI detectors can be considered as one solution. The challenges, however, is that the sophisticated nature of models make it extremely difficult to detect whether AI was involved in content generation, and *to what extent* it was involved. By design, Generative AI models gradually evolve and improve on their ability to emulate human language [33]. Prior to the age of generative AI, traditional plagiarism detection tools like Turnitin achieved success through a simple matching between the text submitted by student and text from anywhere else [10]. Unfortunately, it is the same, simple approach fall short for addressing more advanced algorithms like generative AI. A number of groups and organizations have risen to the challenge by taking different approaches for detecting AI-generated content (as will be identified and described in the next section). However, the fragmented nature of these endeavors lack coordination to make the collective progress toward innovative solutions. While ongoing efforts alternatively involve human oversight as the protective shield [27], questions remain about how generative AI has already seeped in the realm of research, and how we can leverage the state-of-the-art in ways that are ethical and effective.

[1]`https://www.springeropen.com/get-published/editorial-policies`

[2]`https://2023.aclweb.org/blog/ACL-2023-policy`

3. METHODS

In this study, we aim to examine how well current methods of AI content detection would be able to identify generated content aligned with the Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK) communities. These two research communities represent the most prominent publication venues for data–driven research in education, as evidenced by their similarities [14]. Using the most recent proceedings of these two conferences, we aim to conduct an analysis to compare the predictive performance of current popular AI content detection tools under several simulated scenarios. Specifically, we build a dataset comprised of the titles and abstracts of all work published in the 2022 proceedings of the EDM [4] and LAK [3] conferences. Using ChatGPT (GPT-4 Turbo), we use the titles to generate new abstracts for each paper and evaluate 5 detection models in their ability to identify the generated content. In addition, we use ChatGPT to generate a revised version of the human-written abstracts to examine how revised human language affects the detection models' performances. Finally, we look at how partially-generated content may also have an effect on the detection models' performances.

3.1 Data Sources

Our methodology centers on ChatGPT, mainly considering the GPT-4 Turbo model. This model currently exhibits stronger capabilities in all aspects compared to the GPT-3.5 Turbo model also available on the ChatGPT platform. Additionally, the model's accessibility on the ChatGPT platform provides practical support for our research; the availability of the GPT models through programmatic interfaces contributed significantly to the feasibility of our study. As confirmed by the comparative analysis done by Borji and Mohammadian, ChatGPT, particularly its GPT-4 Turbo iteration, stands out among its peers in terms of robustness and versatility [11]. There are currently countless models and platforms with powerful text generation capabilities, such as ChatGPT (GPT-4 Turbo and GPT-3.5 Turbo), Google Bard (Gemini), Claude 2.1, Llama 2, and Bing Chat. ChatGPT, based on the GPT model, remains unparalleled in recognition and widespread use in academia and other fields [25]. For instance, Bing Chat's use of the GPT-4 model has bolstered its influence and creative scope in conversational AI [32].

In addition to the performance of ChatGPT over its competitors, the timing of the release of this tool in conjunction with conference deadlines provided a unique opportunity to

Table 1: Composition of Research Dataset

Data Label	Content Source
H	Human-authored
GPT	ChatGPT (GPT-4-Turbo Model)
GPTR	ChatGPT Revision of Human-authored
GPT/H	50% ChatGPT + 50% Human-authored
H/GPT	50% Human-authored + 50% ChatGPT

study this tool. To explore AI's impact in academia and publishing, we carefully selected the proceedings of LAK22: The 12th International Learning Analytics & Knowledge Conference [3] and EDM2022: The 15th International Conference on Educational Data Mining [4] as our primary sources of data. One of the main reasons for choosing these two conference proceedings as the subject of our study is their timeline in relation to ChatGPT's release. The submission deadline for full paper types at LAK22 was October 4, 2021, while the deadline for other types of submissions was January 31, 2022 [2]. For EDM2022, the paper submission deadline was March 6, 2022, and the deadline for other types of submissions was May 8, 2022 [1]. Given that ChatGPT was released after November 30, 2022, the period before this date, although marked by burgeoning AI capabilities, had not yet reached a level where it could make substantial contributions to academic paper authors, nor had it attracted such widespread attention and use [40]. In this way, by observing papers from these conferences within our dataset, 1) there is an increased likelihood that the paper abstracts were written by humans without significant input from generative AI, and 2) the release of the conference proceedings is unlikely to be included as training to the versions of GPT released to the public at the time of conducting our analyses. Overall, the LAK22 and EDM2022 conference proceedings included a total of 123 and 118 works, respectively.

After identifying the source of the original research datasets, we began to browse the content. While our research revolves around the nexus of ChatGPT's text generation and its implications in academia, it is paramount to define the bounds of our investigation. At the time of our study, ChatGPT was a monomodal system, with text being its primary forte. This posed certain challenges: mainly, text in research papers is often intertwined with diverse media, such as diagrams, images, or even supplemental links and videos.

To avoid potential interference from media complexity present in conference papers, our investigation focuses on titles and abstracts. Specifically, we collected the titles and abstracts of all submissions in the LAK22 and EDM2022 proceedings, forming the human-authored segment of our dataset. These elements encapsulate the essence of academic works, making them highly suitable for our purpose. Moreover, our focus on the unimodal nature of this data provides opportunity for future research to continue in this area. This is especially relevant considering ChatGPT's recent advancements in handling multiple modalities [29].

3.2 Dataset Generation

ChatGPT, which encompasses both the GPT-3.5 Turbo and GPT-4 Turbo models, has marked its dominance in the AI text-generation domain. Given the nuanced disparities in text-generation capabilities between GPT-3.5 Turbo and GPT-4 Turbo, as corroborated by Zhan et al. [41], our research solely harnesses the GPT-4 Turbo model for content generation. We use ChatGPT with the human-written titles and abstracts to expand our data[3] into the following 5 distinct categories (summarized in Table 1):

Human-authored content (H): Original abstracts penned by human, as present in the LAK22 and EDM 2022 conference proceedings. The temporality of these proceedings assures their human origination, negating the influence of advanced AI text generation.

GPT generated content (GPT): Abstracts synthesized purely by ChatGPT. Each research title was presented to ChatGPT with the directive: *"With the following paper title, write a 250-word abstract for a {a learning analytics | an educational data mining} research article: {paper_title}."* An example of this type of content is presented in Figure 1 using the title of this current paper.

GPT revised content (GPTR): Abstracts wholly generated by ChatGPT, albeit rooted in the essence of the original human-authored content (H). The model was given an academic "frame of mind" using a refined prompt: *"I would like you to be a professional educational data mining researcher. Based on the following paper title and abstract, please help me polish the abstract and rewrite it into a 250-word abstract for this research paper. Please only return the revised abstract. Article Title: {paper_title} Abstract: {paper_abstract}."* Such a mode of text generation has become a prevalent practice, especially amongst non-native English speakers, where the foundational thoughts are human-derived, but their expression is reshaped through ChatGPT, often aiding in polishing, embellishing, or translating original ideas [24]. An example of this type of content is presented in Figure 1 using the title and abstract of this current paper.

GPT and Humans Content (GPT/H): A hybrid assembly of abstracts, juxtaposing equal portions of GPT content (50%) and human-authored (50%), sequenced with the GPT content preceding the human content. This is to understand further how the AI content detector works when part of AI-generated content is in the article.

Humans and GPT Content (H/GPT): In contrast to the presentation order of GPT/H, the first half was human-authored content (50%) then GPT-generated content (50%). This data was further used to explore the impact of different content sequences when the AI detector is working.

3.3 AI Text Content Detector

We devised a detector selection scheme based on three criteria that simulates the real-world use of AI content detectors. The first criterion is the ranking in Google searches [20], the second is the ease of use and user-friendly interface, and the third is the provision of an API for broader application contexts. This study also takes into account the preference for free AI content detectors in underfunded educational settings, including two completely free-to-use detec-

[3] `https://osf.io/6rqvx/?view_only=4624f4c7605348ef9ee0f70d6000d51e`

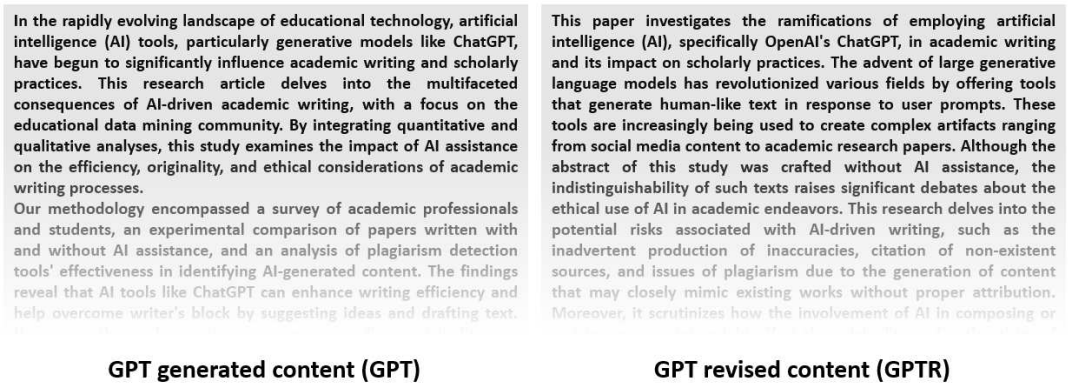


Figure 1: Examples of GPT generated (GPT) and GPT revised (GPTR) abstracts based on the abstract and title of this current paper. This was not included in our analysis and is included here for exemplary purposes.

Table 2: Mean Prediction Scores of AI Content Detectors

	LAK Dataset					EDM Dataset				
Detector	**H**	**GPT**	**GPTR**	**GPT/H**	**H/GPT**	**H**	**GPT**	**GPTR**	**GPT/H**	**H/GPT**
ContentDetector.AI	0.234	0.355	0.152	0.268	0.298	0.246	0.391	0.240	0.317	0.313
ZeroGPT	0.080	0.011	0.021	0.066	0.019	0.022	0.003	0.013	0.011	0.008
GPTZero	0.466	0.502	0.485	0.490	0.501	0.258	0.517	0.505	0.458	0.486
Originality.ai	0.079	0.955	0.941	0.502	0.623	0.119	0.975	0.993	0.609	0.566
Winston.ai	0.069	0.820	0.181	0.426	0.422	0.125	0.759	0.524	0.447	0.474

tors and one with conditional ongoing free use. We evaluated five AI content detectors: ContentDetector.AI, ZeroGPT, GPTZero, Originality.ai, and Winston.ai. Here are the details of these five different detectors: **ContentDetector.AI**[4] is a free detector that recently underwent a version iteration, introducing the more advanced v2 model. This new model boasts enhanced detection capabilities. **ZeroGPT**[5] is a free, open-source project available on GitHub. It provides an easy-to-use interface for AI content detection functions and easy-to-call ports at no cost. **GPTZero**[6] , recognized for its efficacy in academia, journalism, and e-commerce, distinguishes between AI and human-generated content. Its features include a Chrome extension to scan text from the Internet, a Human Writing Report for authenticity validation, and an API to integrate applications. It also integrates with Canvas for educational purposes. GPTZero provides partial free services. However, a monthly subscription is required beyond a certain word limit and number of queries. Its continuous free service availability has led to widespread use [5]. **Originality.ai**[7] operates on a monthly subscription basis. It is built primarily for large web content publishers, aiming to maintain original, AI-free, and plagiarism-free content. Originality.ai claims an average accuracy of 99.41% in detecting texts from these AI models [6]. Its feature set includes AI-generated content detection, plagiarism scanning, website scanning, and Chrome extension integration. **Winston.ai**[8] provides some free credits for trial use, after which it shifts to a monthly subscription model for its services.

[4] https://contentdetector.ai/
[5] https://www.zerogpt.com/
[6] https://gptzero.me/
[7] https://originality.ai/
[8] https://gowinston.ai/

After identifying the AI content detectors that this study focused on, we used these five detectors to test two datasets we created, each containing five different components (H, GPT, GPTR, GPT/H, H/GPT) as listed in Table 1 and obtained probabilistic predictions from each model.

3.4 Evaluation of AI Content Detectors

We evaluate the detector models along metrics of RMSE and AUC as well as an examination of the mean scores produced by the models as a measure of potential bias; for all models, scores range from 0 to 1 with a higher score indicating content likely authored by AI.

RMSE is a standard metric in statistical analysis and machine learning that measures the average magnitude of errors between predicted values [13] (in this case, the scores given by the detectors) and ground truth values (the true nature of the content as human-authored or AI-generated).

For each type of content, the RMSE (Root Mean Square Error) must be calculated. Based on the actual conditions of each different type of content, we can know the ground truth values of different types of content: the ground truth for human-written content (H) is 0, for GPT-generated content (GPT) it is 1, and for GPT-rewritten content (GPTR) it is 1. Lastly, for mixed content (GPT/H and H/GPT), the ground truth value is not so easily identified. One formulation of this mixed content could simply represent these as 1 in alignment to the context of other generated content above; this effectively observes this data as generative if it contains any generated content. Alternatively, since we synthetically ensure that exactly half the content is generated and the other half is human-written, we might expect the

Table 3: Root Mean Square Error (RMSE) of AI Content Detectors

	LAK Dataset					EDM Dataset				
Detector	**H**	**GPT**	**GPTR**	**GPT/H**	**H/GPT**	**H**	**GPT**	**GPTR**	**GPT/H**	**H/GPT**
ContentDetector.AI	0.287	0.664	0.857	0.274	0.243	0.303	0.629	0.776	0.234	0.244
ZeroGPT	0.253	0.993	0.986	0.480	0.492	0.102	0.997	0.992	0.497	0.495
GPTZero	0.474	0.498	0.518	0.048	0.012	0.320	0.483	0.497	0.119	0.074
Originality.ai	0.205	0.172	0.177	0.410	0.387	0.252	0.101	0.038	0.411	0.421
Winston.ai	0.190	0.287	0.870	0.337	0.340	0.254	0.354	0.549	0.242	0.234

Table 4: AUC of AI Content Detectors

	LAK Dataset			EDM Dataset		
Detector	**AUC GPT>H**	**AUC GPT>GPTR**	**Mean GPT>GPTR>H**	**AUC GPT>H**	**AUC GPT>GPTR**	**Mean GPT>GPTR>H**
ContentDetector.AI	0.738	0.863	0.293	0.737	0.850	0.254
ZeroGPT	0.765	0.846	0.133	0.680	0.802	0.271
GPTZero	0.601	0.928	0.268	0.605	0.861	0.585
Originality.ai	0.902	0.698	0.537	1.000	0.581	0.280
Winston.ai	1.000	0.953	0.667	0.991	0.745	0.690

models to ideally produce probabilistic estimates closer to 0.5. While we cannot expect that the models were trained to recognize this type of case, we observe ground truth values at both 0.5 and 1 for this data. In the case of ground truth values at 0.5, we observe the predicted value as correct if the estimate falls within 0.1 of this value (i.e. 0.4-0.6).

In this context, the RMSE calculation adopts a frequently employed approach, where the actual value is determined by considering the ratio of AI-generated text. In this scenario, where the material comprises an equal mix of human and AI-generated content, the actual value is established at 0.5. However, this study also investigates various actual ground truth values for such amalgamated content and employs them as a benchmark for deeper investigation.

The Area Under the Curve (AUC) metric serves as a pivotal analytical tool in our evaluation of AI content detectors, facilitating a nuanced understanding of their performance across various content types. The AUC metric, particularly valuable in the context of Receiver Operating Characteristic (ROC) analysis, provides an aggregate measure of model performance across all possible classification thresholds [21]. In essence, a higher AUC value, approaching 1, signifies a model's superior capability to distinguish between classes—in this case, AI-generated and human-authored content. Conversely, a value closer to 0 indicates a model's tendency to incorrectly classify the content, with a score of 0.5 denoting no discriminative power. The value of including AUC among our metrics is that it does not rely on the identification of rounding thresholds; the AUC value is not sensitive to cases where a model is biased toward a particular label, so long as the ordering of predictions is maintained by the model (i.e. higher predictions are made for positive classes than for negative classes).

The initial focal comparison is between GPT and Human-Authored Content such that it is expected that predictions for generated content should be higher than samples of human-written content (i.e., GPT>H). The primary aspect under consideration is the AUC when differentiating between AI-generated content (GPT) and content created by humans (H).

Secondly, comparisons between GPT and GPT Revised Content should result in cases where higher predictions are made for purely GPT content (GPT>GPTR). Another aspect of our AUC analysis focuses on comparing content generated solely by artificial intelligence (GPT) with content created by humans and then refined or rephrased with the help of GPT (GPTR). This aspect provides valuable insights into the detectors' ability to differentiate between content that is purely AI-generated and content that has been humanized with the assistance of GPT.

Thirdly, we focus on a three-way hierarchical discrimination observing the three iterative levels of generative content. The ultimate goal of this analysis is to assess how well detectors can conform to a hierarchical discrimination framework: content generated by GPT (GPT) should receive a higher score than content enhanced by GPT (GPTR) and created by humans (H), which in turn should be rated higher than content authored solely by humans (GPT>GPTR>H). This hierarchical structure captures the diverse levels of AI involvement in text production, ranging from complete automation (GPT) to partial enhancement (GPTR), and finally to purely human creativity (H).

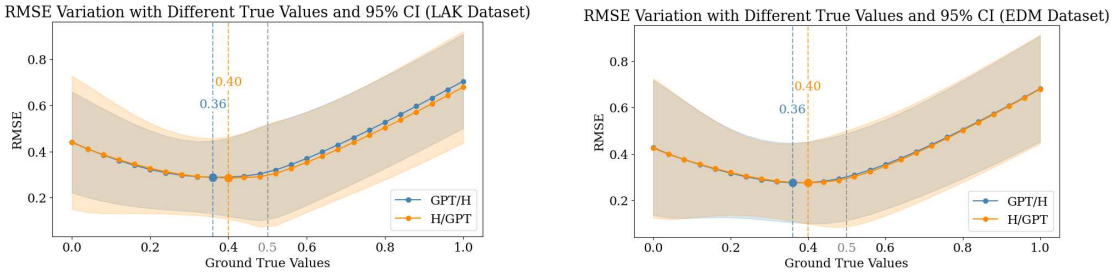
4. RESULTS

The selection and evaluation of AI content detectors, as detailed in the previous section, were designed to reflect real-world applicability and relevance, particularly in educational settings where resources may be limited.

Our analysis of AI content detectors across the LAK and EDM datasets highlights the distinct abilities of these models in identifying human-authored and AI-generated content. These results can be seen across Tables 2. Within these, Originality.ai stands out for its exceptional accuracy in recognizing AI-generated content, as shown by its high overall performance.

Table 5: RMSE Values for GPT/H and H/GPT Content with Different Ground Truth Values

	LAK Dataset				EDM Dataset			
Detector	**GPT/H_0.5**	**H/GPT_0.5**	**GPT/H_1**	**H/GPT_1**	**GPT/H_0.5**	**H/GPT_0.5**	**GPT/H_1**	**H/GPT_1**
ContentDetector.AI	0.274	0.243	0.746	0.715	0.234	0.244	0.699	0.705
ZeroGPT	0.480	0.492	0.957	0.986	0.497	0.495	0.993	0.994
GPTZero	0.048	0.012	0.512	0.499	0.119	0.074	0.554	0.519
Originality.ai	0.410	0.387	0.645	0.525	0.411	0.421	0.557	0.601
Winston.ai	0.337	0.340	0.661	0.666	0.242	0.234	0.602	0.575



(a) (b)

Figure 2: RMSE Variation with Different Ground True Values and 95% CI

Looking at the Mean Prediction Scores, we find that ContentDetector.ai has a good track record for identifying human content but is less consistent with AI-generated text. ZeroGPT, though less effective at identifying AI content, is quite accurate with human text. GPTZero offers a balanced approach to detection, with a slight bias towards flagging content as AI-generated as compared to the other methods. Originality.ai, on the other hand, excels at pinpointing AI-generated content, proving its worth as a premium service. Winston.ai shows reliable detection abilities but struggles somewhat with mixed content types.

The Root Mean Square Error (RMSE) metric reflects the confidence of AI detectors across the datasets, as detailed in Table 3. With a lower RMSE denoting better performance, we found differing levels of precision among the detectors. For human-authored content (H), ContentDetector.AI exhibits an RMSE of 0.287 and 0.303 for the LAK and EDM datasets, respectively. In contrast, ZeroGPT's performance is more varied, with a higher RMSE for AI-generated content, peaking at 0.993 in both datasets, suggesting less consistency in its detection capabilities despite the high confidence of the model.

GPTZero's results are moderate, with its lowest RMSE at 0.048 for mixed content starting with AI-generated text in the LAK dataset. This indicates a closer alignment with the anticipated ground truth in specific mixed content scenarios. Originality.ai, with its lowest RMSE scores of 0.172 and 0.101 for AI-generated content in the LAK and EDM datasets respectively, underscores its proficiency in identifying AI-written text. Winston.ai's RMSE values are relatively low across the board, with its performance on human-authored content (H) demonstrating the least deviation from the ground truth, especially in the EDM dataset, where it scored 0.254. This detector also maintains a balanced performance in discerning mixed content types, with RMSE values that do not exhibit extreme variances

4.1 AUC Analysis

The Area Under the Curve (AUC) metric is instrumental in evaluating the discriminative power of AI content detectors, as encapsulated in Table 4. This statistical tool reflects a model's ability to differentiate between human-authored and AI-generated content, with a value of 1 indicating perfect discrimination and 0.5 implying no better than chance.

The AUC values for the GPT versus Human-Authored content (GPT>H) classification task vary among detectors, with Winston.ai achieving perfect scores (1.000 for the EDM dataset and 0.991 for the LAK dataset), suggesting an impeccable separation between the two content types. Originality.ai also performs well, with an AUC of 0.902 for the LAK dataset, indicating a high level of accuracy in distinguishing between human and GPT-generated content. On the contrary, the AUC for GPT versus GPT revised content (GPT>GPTR) presents a different scenario. GPTZero's AUC of 0.928 for the LAK dataset signifies a robust capability to discern between purely AI-generated abstracts and those that have been revised by GPT, whereas Originality.ai exhibits a lower AUC of 0.698, suggesting room for improvement in this specific task. The mean AUC for hierarchical classification, known as GPT > GPTR > H, reveals that Winston.ai again exhibits higher performance with a mean score of 0.667 for the LAK dataset, closely followed by Originality.ai with a mean score of 0.537. These scores indicate the proficiency of these detectors in recognizing the varying levels of AI involvement in content creation. It should be noted that, while some detectors excel in specific areas, a comprehensive analysis requires a balanced consideration of all AUC values across different classification tasks.

4.2 Evaluating Detector Performance on Mixed Content

Recognizing the inherent challenges in setting appropriate ground truth values for mixed content, we acknowledged that in real academic writing scenarios, a precise value indi-

cating the exact proportion of AI-generated content is often elusive. As previously discussed, the data formats of GPT/H and H/GPT were added to examine this impact. Similarly, as described in a previous section, we are able to reasonably assume a theoretical ground truth value of 0.5, ideally reflecting the proportion of AI-generated content within the overall text. Such a setup provides a basis for evaluating how closely the models can approximate this balance when assessing mixed content.

Table 5 delineates the RMSE values obtained when AI detectors are tasked with evaluating content that embodies an even distribution of human and AI-generated text. The table assesses the performance at two ground truth values, 0.5 and 1, reflecting the balanced nature and the fully AI-generated nature of the content, respectively. The initial thought from the table pointed towards a need for a more detailed investigation, as a clear understanding could not be derived solely from these discrete values. Consequently, we extended our analysis to include a graphical representation, which allowed for a continuous examination of the RMSE as the ground truth values varied between 0 and 1. This approach facilitated a more nuanced observation of the performance trends across the entire spectrum of potential ground truth values.

To extend our analysis of these models in the case of these split data formats, we decided to visualize the performance dynamics of the common AI text detectors. We opted for a graphical representation, plotting dynamic Confidence Interval (CI) graphs to intuitively discern the overall tendencies of these two mixed content types. The visualization aimed to showcase where the models generally position their outputs in relation to the ground truth value. To achieve this, we crafted a figure utilizing a 95% confidence interval to dynamically represent the data. The shaded areas in the graph depict the outcomes across all detectors within our dataset, while the plotted lines represent the average RMSE values for different detectors.

Through this graphical analysis, our objective was to identify the location of the lowest point on the average RMSE line, which would represent the collective behavior of multiple AI detectors in handling mixed content. Pinpointing this minimum on the graph would elucidate the ground truth value at which the majority of models yield their optimal performance, thus shedding light on how these detectors typically respond to mixed texts in terms of their error margins.

This is analyzed through Root Mean Square Error (RMSE) metrics and visualized through Figure 2. These graphs illustrate a discernible pattern where the RMSE values are lower in the middle range, peaking towards the edges—a trend that aligns with our structured dataset comprising an equitable split of human-written and AI-generated text. Optimally, we would anticipate the lowest RMSE values to occur at a ground truth value of 0.5, reflecting an accurate detection of the balanced content. Yet, the data reveals a compelling deviation: both LAK (Figure 2(a)) and EDM (Figure 2(b)) datasets exhibit their nadirs closer to zero, with marked points at 0.36 and 0.40 respectively. This observation indicates a systematic inclination among current AI detectors to classify evenly weighted mixed content as predominantly human-authored.

5. LIMITATIONS AND FUTURE WORK

While our study provides insights into the performance of AI content detectors, it is important to acknowledge the constraints. The range of AI detector tools we examined is not comprehensive. Considering the rapid evolution of AI and detection technologies, our findings offer only a snapshot in time. A more extensive exploration of a variety of tools over a longer period is necessary to gain a comprehensive understanding of the field. This study also focuses on abstracts from two conferences, which may limit the applicability of the results to other disciplines or types of text. Given the significant variations in the style and implementation of AI-generated content, future research should encompass a broader range of academic texts to capture this diversity. Additionally, the assumption that human-generated texts represent an ideal standard is being challenged as AI assistance becomes more integrated into writing processes, often without explicit recognition. This highlights the need for a nuanced approach to defining 'ground truth' in the context of AI-generated content.

6. DISCUSSION

From the AI content detector performance on the LAK and EDM datasets, we see a clear distinction between subscription-based and free detectors in terms of their ability to pinpoint AI-generated content. Originality.ai, a subscription-based service, stands out for its high performance in detecting AI-generated text. On the other hand, freely available tools like ContentDetector.AI and ZeroGPT excel in identifying human-authored content but fall short in recognizing AI-generated material accurately. The analysis using RMSE and AUC metrics also has shed light on the precision and discrimination ability of these detectors. While some detectors excel with specific content types, our results point to the need for an improved uniformity in performance across various content to better serve academic and professional needs. The most important finding this study is the challenge detectors face with mixed content. A noticeable trend was found towards classifying mixed content as mainly human-authored, a pattern consistent across both datasets. Especially noteworthy is the RMSE analysis, where values did not align as expected with a ground truth of 0.5, indicating a bias towards underestimating the AI component in mixed content. This finding reveals a significant area for improvement in detecting models, emphasizing the need for enhanced sensitivity to the text characterizing a mixed contribution.

This study provides a critical examination of AI content detectors in the face of evolving generative AI, offering insights into both the strengths and shortcomings of current detection tools. As the current AI detection tools are still unable to consistently differentiate between text generated by humans and AI, it is not recommended to critique solely based on this factor.

7. ACKNOWLEDGMENTS

We would like to thank NSF (e.g., 2331379, 1903304, 1822830, 1724889), as well as IES (R305B230007), Schmidt Futures, MathNet, and OpenAI.

8. REFERENCES

[1] Edm2022 important dates, 2022. Accessed on 2024-01-20.

[2] Lak22 important dates, 2022. Accessed on 2024-01-20.

[3] *Proceedings of the 12th International Conference on Learning Analytics & Knowledge (LAK22)*, Online, 2022. Society for Learning Analytics Research (SoLAR).

[4] *Proceedings of the 15th International Conference on Educational Data Mining*, Durham, UK and Online, 2022. International Educational Data Mining Society.

[5] Gptzero review 2024: Is this ai detection tool worth it?, 2024. Accessed: Jan 19, 2024.

[6] My honest review of originality.ai - does it actually work? 2024. Accessed: Jan 19, 2024.

[7] S. A. Antu, H. Chen, and C. K. Richards. Using llm (large language model) to improve efficiency in literature review for undergraduate research. 2023.

[8] A. Balat and İ. Bahşi. We asked chatgpt about the co-authorship of artificial intelligence in scientific papers. 2023.

[9] S. Baral, A. Santhanam, A. F. Botelho, A. Gurung, and N. Heffernan. Automated scoring of image-based responses to open-ended mathematics question. In *The Proceedings of the 16th International Conference on Educational Data Mining*, 2023.

[10] T. Batane. Turning to turnitin to fight plagiarism among university students. *Journal of Educational Technology & Society*, 13(2):1–12, 2010.

[11] A. Borji and M. Mohammadian. Battle of the wordsmiths: Comparing chatgpt, gpt-4, claude, and bard. *GPT-4, Claude, and Bard (June 12, 2023)*, 2023.

[12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[13] T. Chai and R. R. Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

[14] G. Chen, V. Rolim, R. F. Mello, and D. Gašević. Let's shine together! a comparative study between learning analytics and educational data mining. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 544–553, 2020.

[15] Y. K. Dwivedi, L. Hughes, E. Ismagilova, G. Aarts, C. Coombs, T. Crick, Y. Duan, R. Dwivedi, J. Edwards, A. Eirug, et al. Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57:101994, 2021.

[16] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, et al. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642, 2023.

[17] T. N. Fitria. Grammarly as ai-powered english writing assistant: Students' alternative for writing english. *Metathesis: Journal of English Language, Literature, and Teaching*, 5(1):65–78, 2021.

[18] T. N. Fitria. Quillbot as an online tool: Students' alternative in paraphrasing and rewriting of english writing. *Englisia: Journal of Language, Education, and Humanities*, 9(1):183–196, 2021.

[19] C. A. Gao, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, and A. T. Pearson. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, 6(1):75, 2023.

[20] Google. Search results for "ai detector". `https://www.google.com`, 2023. Accessed: 2023-09-20.

[21] J. Huang and C. X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[22] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[23] J. Leinonen, P. Denny, S. MacNeil, S. Sarsa, S. Bernstein, J. Kim, A. Tran, and A. Hellas. Comparing code explanations created by students and large language models. *arXiv preprint arXiv:2304.03938*, 2023.

[24] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou. Gpt detectors are biased against non-native english writers. *arXiv preprint arXiv:2304.02819*, 2023.

[25] Z. Liu, T. Zhong, Y. Li, Y. Zhang, Y. Pan, Z. Zhao, P. Dong, C. Cao, Y. Liu, P. Shu, et al. Evaluating large language models for radiology natural language processing. *arXiv preprint arXiv:2307.13693*, 2023.

[26] K. Martin. Ethical implications and accountability of algorithms. *Journal of business ethics*, 160:835–850, 2019.

[27] B. Meskó and E. J. Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120, 2023.

[28] A. Moreno and T. Redondo. Text analytics: the convergence of big data and artificial intelligence. *IJIMAI*, 3(6):57–64, 2016.

[29] OpenAI. Chatgpt can now see, hear, and speak, 2023.

[30] S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer. The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590*, 2023.

[31] M. R. Rahman, R. S. Koka, S. K. Shah, T. Solorio, and J. Subhlok. Enhancing lecture video navigation with ai generated summaries. *Education and Information Technologies*, pages 1–24, 2023.

[32] J. Rudolph, S. Tan, and S. Tan. War of the chatbots: Bard, bing chat, chatgpt, ernie and beyond. the new ai gold rush and its impact on higher education. *Journal of Applied Learning and Teaching*, 6(1), 2023.

[33] V. S. Sadasivan, A. Kumar, S. Balasubramanian,

W. Wang, and S. Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

[34] H. Salehi and R. Burgueño. Emerging artificial intelligence methods in structural engineering. *Engineering structures*, 171:170–189, 2018.

[35] A. Shibani, R. Rajalakshmi, F. Mattins, S. Selvaraj, and S. Knight. Visual representation of co-authorship with gpt-3: Studying human-machine interaction for effective writing. In *Proceedings of the 16th International Conference on Educational Data Mining*. International Educational Data Mining Society, ERIC, 2023.

[36] A. M. Turing. *Computing machinery and intelligence*. Springer, 2009.

[37] E. A. Van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023.

[38] R. S. Wallace. *The anatomy of ALICE*. Springer, 2009.

[39] Z. Wang, Y. Guo, T. Xia, B. Ye, and P. Kar. Detection of covid-19 through thermal and voice sensing using smartphone. In *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*, pages 2–7. IEEE, 2022.

[40] H. Yu. Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14:1181712, 2023.

[41] H. Zhan, X. He, Q. Xu, Y. Wu, and P. Stenetorp. G3detector: General gpt-generated text detector. *arXiv preprint arXiv:2305.12680*, 2023.